17. Reading external data

Pandas provides several methods to read external data into a DataFrame.

1. Read from a CSV file

```
pd.read_csv()
```

Parameters,

- `sep` : specify a delimiter.
- `header` : specify the row to use as column names.
- `names` : specify custom column names.
- `dtype` : specify data types for the columns.
- `parse_dates` : to parse dates.
  Ex:

```
df = pd.read_csv('data.csv', sep=',', header=0, dtype={'column1': int},
parse_dates=['date_column'])
```

2. Read from an excel

```
pd.read_excell()
```

Parameters,

- `sheet_name` : specify which sheet to read
  Default to first sheet
  If want to read a specific sheet, provide it's name to this parameter.
  If want to read all sheet at single go then pass `None` for this parameter.
- `usecols` : specify which columns to read
- `skiprows` : skip rows at the beginning of the file
- `nrows` : read a specific number of rows
  Ex:

```
df = pd.read_excel('data.xlsx', sheet_name='Sheet1', usecols='A:C',
skiprows=1, nrows=100)
```

3. If you want to read and write DataFrames faster for internal purpose use **pickle file**

formats as the reading time is much faster then excel or CSVs.

```
df.to_pickle(path)
#or
df = pd.read_pickle(path)
```

`to_pickle`

used to serialize and save a DataFrame to a pickle file.

```
DataFrame.to_pickle(path, compression='infer', protocol=5,
storage_options=None)
```

- `path` : The file path where the DataFrame will be saved.
- `compression` : (Optional) Specifies the compression mode.
  Ex: `'infer', 'gzip', 'bz2', 'zip', 'xz', 'zstd'`
  Defaults to 'infer'
- `protocol` : (Optional) Specifies the pickle protocol version to be used.
  Default is 5.
- `storage_options` : (Optional) Extra options related to the storage backend.

`read_pickle`

used to deserialize and load a DataFrame from a pickle file.

```
DataFrame.read_pickle(path, compression='infer', storage_options=None)
```

- `path` : The file path from which the DataFrame will be loaded.
- `compression` : (Optional) Specifies the compression mode.
  Ex: `'infer', 'gzip', 'bz2', 'zip', 'xz', 'zstd'`
  Defaults to 'infer'
- `storage_options` : (Optional) Extra options related to the storage backend.

Pandas support read from other data types as well,
Ex:
`pd.read_stata(myfile.dta)`
`pd.read_sas(myfile.sas7bdat)`
`pd.read_hdf(myfile.h5,'df')`
`pd.read_sql()`
`pd.read_json()`
`pd.read_html()`

`pd.read_sql_table()`

`pd.read_gbq()` - Google BigQuery